Non-Speech Aural Communication for Robots

Frederick Heckel and William D. Smart
Department of Computer Science and Engineering
Washington University in St. Louis
Campus Box 1045
One Brookings Drive
St. Louis, MO 63130
United States
{fwph,wds}@cse.wustl.edu

Abstract

Synthesized or pre-recorded speech is often used for communication from robots to humans. In many situations, it may be unnecessary to use speech since much of the information which robots must communicate is often very simple and context-sensitive. In this paper we hypothesize that non-speech aural communication may be more effective in many situations, and we present an experiment to test the effectiveness of non-verbal communication.

Introduction

Synthesized or pre-recorded speech is often used for communication from robots to humans, with varying levels of success. Synthesized speech has improved a great deal in recent years, but is still very unnatural and difficult to generate effectively. Pre-recorded speech can be expensive (in time) to prepare and results in a limited set of expressions which can be used. Pre-recorded phrases can be extended by splicing together individual words, but this destroys the nuance of speech, resulting in a very unnatural voice. Both of these methods suffer from an additional problem: if the vocabulary is too complex, and the robot's speech comprehension abilities are not at a comparable level to its speech generation abilities, users may become very frustrated with the robot's inability to understand spoken commands. In addition, human language is very complex while the state information that a robot needs to communicate is often very simple (ie, "I'm stuck", "I cannot complete this task", "Please assist", "I have achieved my goal", "Please move out of the way", etc).

We believe that, while it may be difficult to communicate complex ideas, non-speech aural cues can be more effective than speech cues for communicating simple state information and requests. Also, non-speech cues may be substantially more pleasing to the ear than the synthesized speech often used in robotic systems and can even be presented as continuous "music". In this paper, we review some of the human-computer interaction literature relating to multimedia interaction, including auditory icons (natural sounds from the world) and "earcons" (generated, abstract sounds). We present an experiment based on recent work in human-robot interaction and methods from human factors engineering.

Copyright © 2006, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

Background

The field of Human-Computer Interaction (HCI) has generated a great deal of literature on different modes of control and feedback in interfaces. The most common mode is the standard visual display with keyboard/mouse input. As computers have increased in processing power, multimedia communication has become an increasingly important area of research (Rosenfeld, Olsen, & Rudnicky 2001; Murayama, Constantinou, & Omata 2004; Perzanowski *et al.* 2001).

The problems of interaction with robots are similar in many ways to the issues faced by HCI researchers. Each field attempts to convey a great deal of information through simple interfaces for interaction. Human-Robot Interaction (HRI) faces additional challenges beyond the traditional HCI domain. For teleoperation of robotic systems, situational awareness becomes a serious problem, as the operator must be able to understand and reason about a rich array of sensory input. Robot sensors are different from traditional human senses, and it can be difficult to create easily-understandable representations of even simple range sensors. Even the modes we are familiar with are often in very different configurations from human expectation: visual sensors may be mounted in surprisingly (to the human operator) high or low positions while aural sensors may have a limited response and the data in both cases may be out of sync with the control signal due to network time delays. The fact that robots have a physical presence in the world makes the field of HRI fundamentally different from HCI, with different questions and potentially very different answers.

Autonomous robots present even more problems, since we cannot assume that the human with whom they will interact will have any particular interface devices available. This often means that all interaction must be in a natural human mode (most commonly aural and visual), or through interfaces carried on the robot. In addition, while computer displays can be used, they may cause a prohibitive power drain on the system. These factors make multimedia interaction paradigms not only attractive for autonomous mobile robots, but actually necessary for effective interaction in many set-

tings.

One of the most commonly used interaction methods is speech. Speech recognition is a difficult task, but is now feasible. It has even found deployment in certain limited-domain commercial systems which we use every day such as automated support lines. Natural language processing (NLP) is an extremely difficult task (and far from solved), but NLP research has generated a great body of work which allows us to interpret moderately complex speech. Methods from NLP are already used successfully in commercial applications, but robots face problems interpreting speech due to self-generated noise (from actuators, fans, or other on-board sources), or unpredictable environmental sources. In addition, robots often have limited processing capabilities due to the high power consumption of general purpose processors. Because the batteries which are used to power robots have limited capacity, robots often cannot support computers with sufficient processing power to perform extensive real-time processing of sensor information. Due to these problems, speech, while valuable, is often not ideal for all situations.

Generating speech for output can also be problematic. Pre-recorded speech is sometimes used for high quality speech output, but synthesized speech is more common because of its greater flexibility. Recording large vocabularies requires a great deal of time, but synthesized speech is often unpleasant to listen to. The HCI literature reveals an additional mode of aural interaction: the "auditory icon" (Gaver 1991) and the "earcon"(Blattner, Sumikawa, & Greenberg 1989). Auditory icons are natural sounds which are common in the real world, while earcons are abstract sounds. These aural cues possess the advantages of being very simple (and so require very little human processing to understand), and auditory icons specifically exploit pre-existing knowledge of certain sounds.

The use of speech can also generate an expectation of a level of intelligence that the robot does not possess. When the robot speaks using complex language, and then does not respond to similarly complex speech, people often become frustrated. We have seen this effect in previous research with robots in social situations; many people will briefly attempt to get the robot's attention using gestures or speech but then become bored and walk away when it does not respond appropriately (Smart *et al.* 2003). In some cases, people actually become angry with the robot and will occasionally retaliate by pressing its emergency stop button. Given all of these factors, we believe that for many situations (especially for mobile service robots), non-speech aural cues may be more effective than language.

There is previous work (Warren 2002) which indicates that non-speech cues are less successful than speech in adaptable autonomy (AA)[1] situations, but AA results are applicable primarily for teleoperation. Since teleoperation studies focus on the operator of the robot, rather than third parties which are interacting physically with the robot, the effectiveness of different interaction modes may be very different. Non-aural, non-verbal communication (gestures, facial expressions, and gaze direction) has been found to be successful in a learning situation (Breazeal *et al.* 2005). To the best of our knowledge, while there has been some evaluation of non-speech cues for communicating state information (Johannsen 2002) for autonomous robots, there have been no rigorous studies on the effectiveness of sounds effects compared to speech in HRI with autonomous robots.

[1]Adaptable autonomy makes use of a sliding scale for robot control. On one extreme is complete operator control of all robot functions, while the other extreme is complete autonomy with no operator intervention.

Figure 1: The experimental platform: Lewis, an iRobot B21r research robot

Preliminary Research

We developed an experiment based on the AAAI Scavenger Hunt (Oh & Dodds 2006).The goal of the scavenger hunt is for the robot to identify, locate, map, and manipulate a number of small objects placed throughout the competition area. The competition area is a dynamic space not closed off to other conference activity. Scoring is based on autonomy, the degree to which the environment must be modified by the robot team, human interaction, accuracy of the environmental map, the ability of the robot to identify objects in different physical locations (on the floor, on a table, high in the air), and time required to complete the task. Robots may make use of human interaction to help in the task. This competition presents a number of challenges which are directly relevant to many more serious applications, such as courier robots.

Our platform is an iRobot B21r research robot fitted with stereo vision on a pan-tilt actuator, an LCD touch-screen monitor, a SICK laser range-finder, and a standard PC sound

card. For this experiment, the robot was teleoperated rather than autonomous. Subjects were not advised that the robot was teleoperated, and the operator controlled the robot from a separate room. Due to technical difficulties, our initial experiments were conducted in the lab rather than at the conference. Doing so eliminated the initial challenge of getting the subject's attention, and allowed for a more controlled experimental environment.

Metrics

Scholtz (2003) describes several roles which humans may fill when interacting with robots: supervisor, operator, mechanic, peer, and bystander. This list is adapted from the HCI literature. The supervisor role is filled by an individual who monitors the overall situation and can make changes to the plans of the robot or immediate actions. Operators are more restricted in their control, dealing with specific action level problems. The mechanic fixes problems at the hardware and software level, but does not make activate different plans or actions during normal operation.

The roles we are most interested in are those of *peer* and *bystander*. The peer role is filled by a user who is working with the robot to perform a task (the peer shares a goal and may have some prior knowledge of the system), while the bystander may have no information about the task at hand and no familiarity with the robot. Scholtz limits the bystander to actions such as stepping in front of the robot to force it to stop or modify its course, but in the scavenger hunt scenario, part of the challenge is to persuade the bystander to become involved in the task and take on the role of peer.

To this end, the metrics we use are those of task performance, awareness of robot state, and aesthetic appeal. The performance of the system can be measured by the amount of time taken to perform the task (find the sought-for objects), and the number of mistakes made (locating the wrong object) during the course of the interaction. State awareness can be best determined via human analysis of trial video while aesthetic appeal can be determined by questionnaire. Our use of questionnaires is limited, due to the questionable reliability and accuracy of questionnaire methods (Stanton *et al.* 2005).

Performance The performance of our system is measured against a baseline of how long it takes a human/robot team to complete the task without any aural signals. Successful interaction will shorten the time necessary to locate all of the objects. For the full scenario with multiple bystanders, success would also be gauged by the robot's ability to engage bystanders.

Awareness of Robot State One of the most important aspects of communication is comprehension of robot state: the current request of the robot, its understanding of communication from the subject, and awareness of any problems (simulated or otherwise) that may occur during the course of the interaction. This is somewhat different from the usual definition of situational awareness used in machine interface studies, but is a similar concept that may be possible to evaluate using methods for measuring situational awareness.

Aesthetic Appeal Aesthetic appeal is the most subjective of the three criteria we are interested in. A more attractive mode of interaction will allow a robot to engage people more easily. For example, a very high-pitched, loud, continuous sound, while potentially quite communicative of a problem, will tend to discourage interaction with the robot (except to find the off button!). Similarly, some auditory icons may be much more appealing than harsh synthesized speech or highly repetitive pre-recorded speech.

Experimental Setup

To experimentally test these ideas, we designed a simple user study to compare the different modes of aural communication. We broke the human-assisted scavenger hunt task down to a small number of (interaction) steps:

- Locate a bystander, approach and begin interaction.
- Through use of aural cues and the LCD monitor, request assistance in locating a specific object.
- Follow the human assistant to the object's location, alerting her if she is moving too quickly.
- Indicate whether the object found is the correct one.
- Request the subject to pick up the object and follow the robot to return it to home base.

These specify certain important interaction types, with different challenge levels. The simplest interactions are indicating if the correct object has been found, and next indicating state during the following phase. Requesting the subject to move the object is next step in difficulty; while this is a fairly complex request, the subject is already committed to helping the robot. Engaging the bystander is slightly more difficult, as it asks the subject to make a commitment of time, and interrupt any current tasks or social engagement. We place requesting assistance to locate the object as the most difficult task because it not only asks further engagement from the subject and a significant investment of attention, but the request "please lead me to this obstacle" is more complex than any of the previous requests, and is difficult to code using auditory icons for the untrained bystander[2].

Our initial experimental work used six different recordings of sheepdog vocalizations. No speech was used, instead, control trials with no sound were used to test our assumption that the sound would provide important cues for the task, and images on the screen alone would not suffice. Subjects were informed that the robot had a habit of losing his toys, and sometimes needed help finding them again. At this point, the control subjects began the interaction. Experimental subjects were first asked to give one or two words

[2] An important aspect of these studies is that we are most interested in the untrained bystander. We expect that with a short period of training, synthesized sounds could be as effective as natural sounds or speech, but they are unlikely to generate the correct response to the completely untrained user

Table 1: Experimental Sounds and Purpose

Sheepdog Vocalizations	
Sound	Intended Interpretation
Double Bark	Attention
Several Barks	Slow down / Emphatic attention
Short Yip	Yes / Confirm
Whimper	Please help
Low to high pitch bark	No / Incorrect / Confused
Several barks (high pitched)	Thanks for the help

describing the dog vocalizations which were being used, and asked whether they were dog owners. After the interaction, these subjects were again asked to interpret the sounds. Each subject helped the robot locate three different objects, and place them in specific locations. Each object was shown on the screen as it was requested. Only simple motor actuation was used with the robot, such as moving the robot through the environment and turning the pan tilt unit to look at the location where the object was to be placed.

Results and Further Experiments

The results of our pilot trials, with three control and three experimental subjects, show that our assumption about the image display was incorrect. The images of the object shown on the small LCD touchscreen dominated the interaction. While we cannot make any strong claims based on these six trials, we can say that in this case, the sound did not have a noticeable effect on trial time or accuracy. Furthermore, any training effect from sound repetition was eclipsed by the training effect of the task itself– because each subtask was very similar (except that objects were to be placed in different locations), the subjects understood clearly what was expected after the first object was located.

Due to these shortcomings, we plan to redesign the experiment to better test our hypothesis. In addition to the dominance of the touchscreen, subjects became annoyed due to the repetitive nature of the task, and subjects spent more time looking for objects than interacting with the robot.

Table 2: Performance (Timing) Results

Time to complete task		
Group	Average Time	Standard Deviation
Sound	6 minutes	0.6 minutes
Control	6 minutes	1.3 minutes

Evaluating the effectiveness of non-verbal and verbal communication with a mobile robot is a more difficult task than we first thought. The task must be simple enough that all communication can be achieved through sound, but variable enough that the exact task cannot be learned after the first phase (while still allowing sufficient repetition to allow any training effect to develop). It must be possible to isolate the effect of the sound next to other communication channels– motion and actuation, for example, can provide a great deal of information in a subtle, hidden channel. Furthermore, even the small amount of data gathered through direct questioning in this experiment was unreliable. At the end of the trial, each subject could not remember hearing at least half the sounds, or expressed uncertainty as to any interpretation, even when they had clearly responded to a sound during the interaction.

A possible modification to the experiment would be for the robot to guide the task, rather than the person. After getting the subject's attention, the robot would lead the subject to a group of objects, and use sounds to indicate which object should be chosen. A simple coding scheme based on color or shape could be developed, and combined with basic yes/no sounds. The subject would choose different objects, and feedback from the robot would help the person decide which was the correct object. The robot could then lead the subject to the second location, and so on. For the sake of data gathering, it would be better to have the subject describe verbally his interpretation of the current intention of the robot, and the meaning of the sounds, as the trial proceeded.

Before running a trial intended to test effectiveness for untrained subjects, an initial study to determine how people interpret the intended sounds is important. Only sounds which have similar interpretations across different groups of listeners should be used; the full experiment with robots is not the time to discover that interpretations vary widely. While our initial experiment was not successful along the axis we intended, it has provided a useful experience for designing future experiments which we hope other researchers can learn from.

Current and Future Directions

While our initial work focuses on using everyday sounds to communicate state information and elicit responses, it suffers from the same problem as pre-recorded speech: the robot will likely sound very repetitive and will have little flexibility in communication. A greater vocabulary of sounds will reduce this problem, but a more elegant solution would be to identify categories of sounds which can communicate the same information. It may be possible to develop a "sound space"– effectively a simple language based on human cognition– that focuses on using changes in pitch and volume. We are currently working with a colleague in psychology to develop this idea further.

Before we can proceed with this work, we must first develop a better framework for quantifying the different communication channels present in the interaction. This may include movement and even social situation along with visual and aural information. Without a better framework for classifying external factors, it will be very difficult to design an experiment to evaluate the effectiveness of non-verbal aural communication.

On the larger scale, communication from robots to humans is a comparatively simple task next to comprehension

of human commands to robots. If a relatively simple language for communicating information from robots to humans can be developed, then it may be possible to reverse it. Changes in pitch and volume are easier to process computationally than a full human language, even given perfect speech recognition capabilities. It may be that such a sound space can be used for two-way human-robot or robot-robot communication.

Further analysis of our initial work will be available at the Fall Symposium.

References

Blattner, M. M.; Sumikawa, D. S.; and Greenberg, R. M. 1989. Earcons and icons: Their structure and common design principles. In *Human Computer Interaction*, 11–44.

Breazeal, C.; Kidd, C. D.; Thomaz, A. L.; Hoffman, G.; and Berlin, M. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Intelligent Robots and Systems*.

Gaver, W. W. 1991. Sound support for collaboration. In *Second European Conference on Computer-Supported Collaborative Work*.

Johannsen, G. 2002. Auditory display of directions and states for mobile systems. In *International Conference on Auditory Display*.

Murayama, Y.; Constantinou, C. E.; and Omata, S. 2004. Intuitive micromanipulation with haptic audio feedback. In *Conference on Computer and Information Technology*, 907–911. Los Alamitos, CA, USA: IEEE Computer Society.

Oh, P., and Dodds, Z. 2006. Scavenger hunt event. http://www.cs.hmc.edu/∼dodds/aaai06/. AAAI 2006. Visited May 1, 2006.

Perzanowski, D.; Schultz, A. C.; Adams, W.; Marsh, E.; and Bugajska, M. 2001. Building a multimodal human-robot interface. *IEEE Intelligent Systems* 16(1).

Rosenfeld, R.; Olsen, D.; and Rudnicky, A. 2001. Universal speech interfaces. *Interactions* 8(6).

Scholtz, J. 2003. Theory and evaluation of human robot interactions. In *Hawaii International Conference On System Sciences*, 125.

Smart, W. D.; Grimm, C. M.; Dixon, M.; and Byers, Z. 2003. (not) Interacting with a robot photographer. In Kortenkamp, D., and Freed, M., eds., *Human Interaction with Autonomous Systems in Complex Environments: Papers from the 2003 AAAI Spring Symposium*, 181–186. Available in AAAI Technical Report WS-03-04.

Stanton, N. A.; Salmon, P. M.; Walker, G. H.; Baber, C.; and Jenkins, D. P. 2005. *Human Factors Methods: A Practical Guide for Engineering and Design*. Ashgate.

Warren, H. L. 2002. Auditory cueing effects on human performance with an adaptive system. Master's thesis, North Carolina State University.